CST0
COMPUTER SCIENCE TRIPOS Part IA

Friday 7 June 2024 13:30 to 16:30

COMPUTER SCIENCE Paper 3

*Answer **one** question from each of Sections A, B and C, and **two** questions from Section D.*

*Submit the answers in five **separate** bundles, each with its own cover sheet. On each cover sheet, write the numbers of **all** attempted questions, and circle the number of the question attached.*

You may not start to read the questions printed on the subsequent pages of this question paper until instructed that you may do so by the Invigilator

STATIONERY REQUIREMENTS
Script paper
Blue cover sheets
Tags

SPECIAL REQUIREMENTS
Approved calculator permitted

SECTION A

1 Databases

(a) A relational algebra is defined over sets of tuples. Explain whether the relational union of two sets (relations) requires the schemas to share attribute names. Does the same consideration apply for intersection? [4 marks]

(b) In the same relational algebra, sets with P and Q records (table lengths) are joined by a binary operator. What is the minimum and maximum number of records in the answer if the operation is union? What if it is a natural join? [4 marks]

(c) A relational database of text books holds a small amount of information about each chapter of each book. Two relations (tables) are used with a total of seven distinct attribute names (fields). Draw a suitable E/R diagram and define the two corresponding relational schemas. Say what forms of key are present. [7 marks]

(d) The textbooks all relate to a common subject area and a more detailed database is now required, perhaps storing the books themselves. What new operations might be wanted from this new system, how would you structure it and would there be benefits from enforcing consistency rules? [5 marks]

2 Databases

(*a*) It is reckoned to be impossible to simultaneously provide the CAP trio in DBMS design: these are C---------y, A----------y and P-------- t-------e. Complete and define these terms and explain why. [4 marks]

(*b*) What is one significant difference between a reflexive relation in discrete maths and a database relation between a domain and itself? Give simple examples of both. Can either form of relation usefully have a one-to-one cardinality? [6 marks]

(*c*) An rDMBS holds tables with these four schemas:

R1:(<u>A, B</u>, C, D), R2:(<u>A, B</u>, C), R3:(<u>A, B</u>, D, E, F) and R4:(<u>A</u>, D).

You are told that values of F are always predictable from values of E, but it might be costly to make that prediction. Also, database updates might be much rarer than reads. What rearrangment of the schemas might be good and why? [5 marks]

(*d*) An `is_a` relation between entity types is defining a two-level hierarchy: for instance, lions and whales are both mammals. There are three basically different ways that a two-level `is_a` hierarchy can be modelled using relational tables. What are they? Multi-level `is_a` relationships commonly arise: which way then might be best? [5 marks]

SECTION B

3 Introduction to Graphics

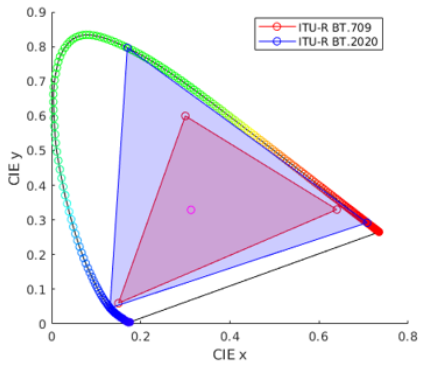
(a) Select the most suitable colour space for the following applications. Justify your choice.

(i) Color picker in painting software. [2 marks]

(ii) An error metric for evaluating colour distortions introduced by a camera. [2 marks]

(iii) Storing high dynamic range textures for real-time rendering. [3 marks]

(b) The diagram below shows a projection of the colour gamut and the primaries of two popular RGB colour spaces.



(i) BT.709 primaries are rather far from the boundary of the visible colour gamut. What motivated that choice of primaries in BT.709? [3 marks]

(ii) Most modern displays do not reproduce all colours from the BT.2020 colour gamut. What motivated that choice of the colour gamut? [3 marks]

(c) You design a colour management system for a new display with atypical primary colours. You measured the spectra of its three primaries and stored measured samples in an $N\times3$ matrix D, in which the columns correspond to the three primaries and the rows to the spectral samples. Find a transformation matrix from BT.709 RGB to the values of the three native primaries of the display that provide a metameric match. You have an $N\times3$ matrix S_{XYZ} with the XYZ colour matching functions and a 3×3 matrix $M_{709\rightarrow\mathrm{XYZ}}$ to transform from RGB BT.709 to XYZ. Both input and output colour spaces are linear. You can ignore normalization constants. [7 marks]

4 Introduction to Graphics

Your task is to design a rendering pipeline for a next-generation augmented reality headset.

(*a*) Explain how you could improve the quality of rendering given the following hardware limitations. Keep your answers short, up to 100 words.

(*i*) The high refresh rate display will be controlled by 6 instead of 8 bits for each colour channel. [3 marks]

(*ii*) The display resolution will be 1024×1024, resulting in a relatively large pixel size when magnified by the lens of the headset [4 marks]

(*iii*) The headset is equipped with an optical see-through display in which the light from the environment is mixed (in an additive manner) with the light from the display. The headset has a sensor that can measure the overall amount of environment luminance reaching the display. [5 marks]

(*b*) The tracking sensors provide the position of the estimated centre of the left eye, e_{L}, in the units of millimetres and the view direction, $\hat{v}$, both in the world coordinates. The distance between the two eyes in millimetres is d_{IOD}. The headset uses a right-handed coordinate system, and the *up* vector is $u = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$. Derive view matrices for the left and the right eye cameras. Illustrate your answer with a diagram. [8 marks]

SECTION C

5 Interaction Design

ChatGPT is a natural language processing tool, developed by OpenAI, driven by AI technology that allows you to have human-like conversations with a chatbot. It has a Free ChatGPT Unlimited version, whose main screen is shown in the figure above, that is a free-to-use AI system. Based on a large language model, ChatGPT can answer questions and assist its users with tasks, such as composing emails and code, and enable users to refine and steer a conversation towards a desired length, format, style, level of detail, and language.

(*a*) Identify four (different) groups of stakeholders of the Free ChatGPT Unlimited tool and provide a short description as to why they constitute a stakeholder group for this tool.

[4 marks]

(*b*) Nielsen's heuristics used for Heuristic Evaluation are: (1) visibility of system status, (2) match between system and real world, (3) user control and freedom, (4) consistency and standards, (5) error prevention, (6) recognition rather than recall, (7) flexibility and efficiency of use, (8) aesthetic and minimalist design, (9) help users recognize and recover from errors, and (10) help and documentation.

Using Figures 1-6, undertake a Heuristic Evaluation of the Free ChatGPT Unlimited.

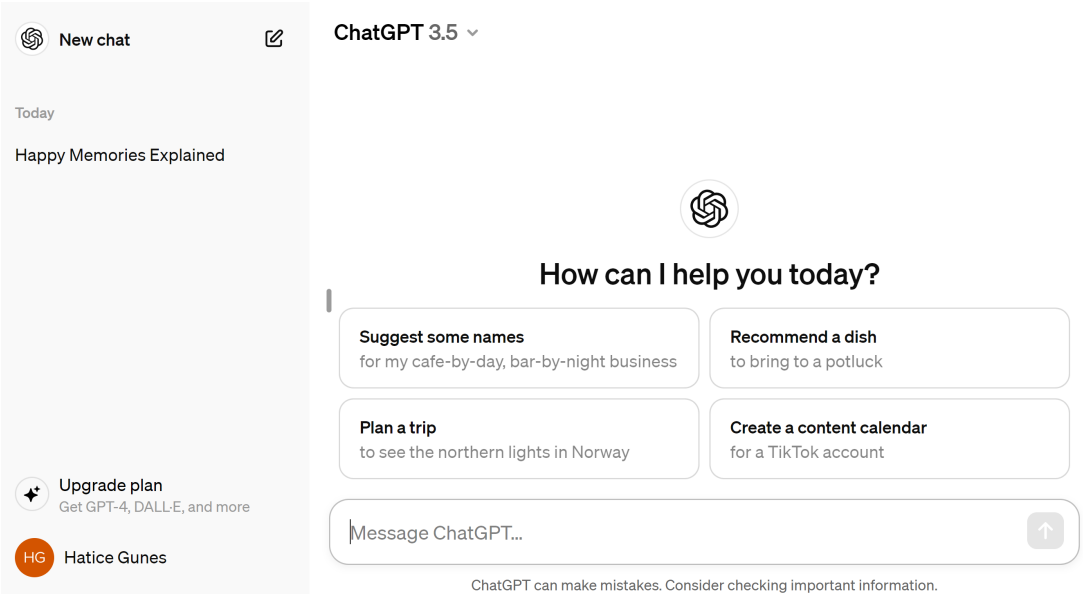


Fig. 1. First / Main screen

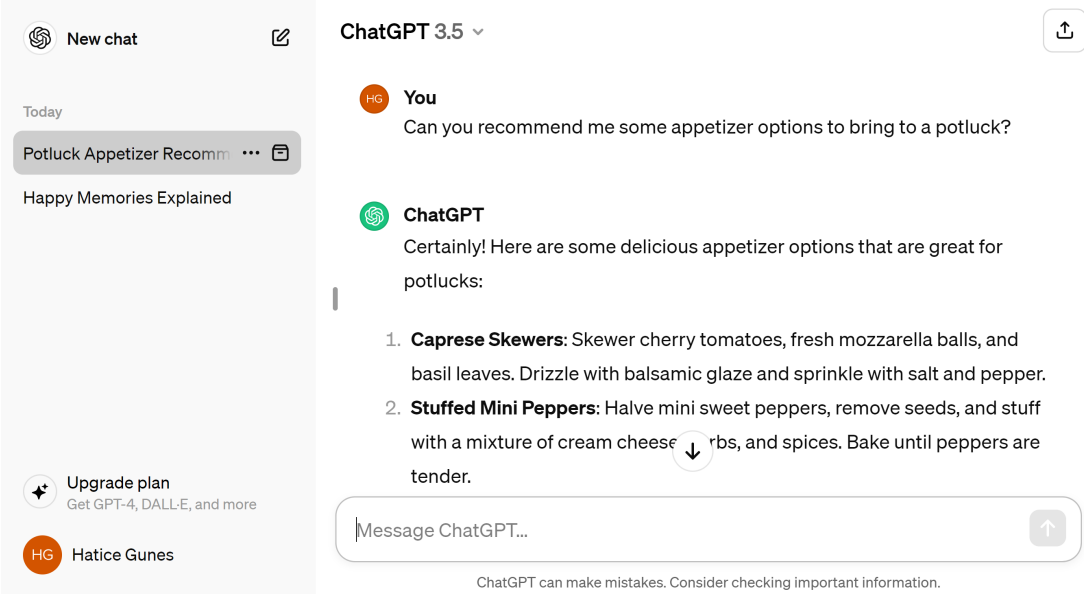


Fig. 2. (starting from the main screen) Asking for appetizer options to bring to a potluck

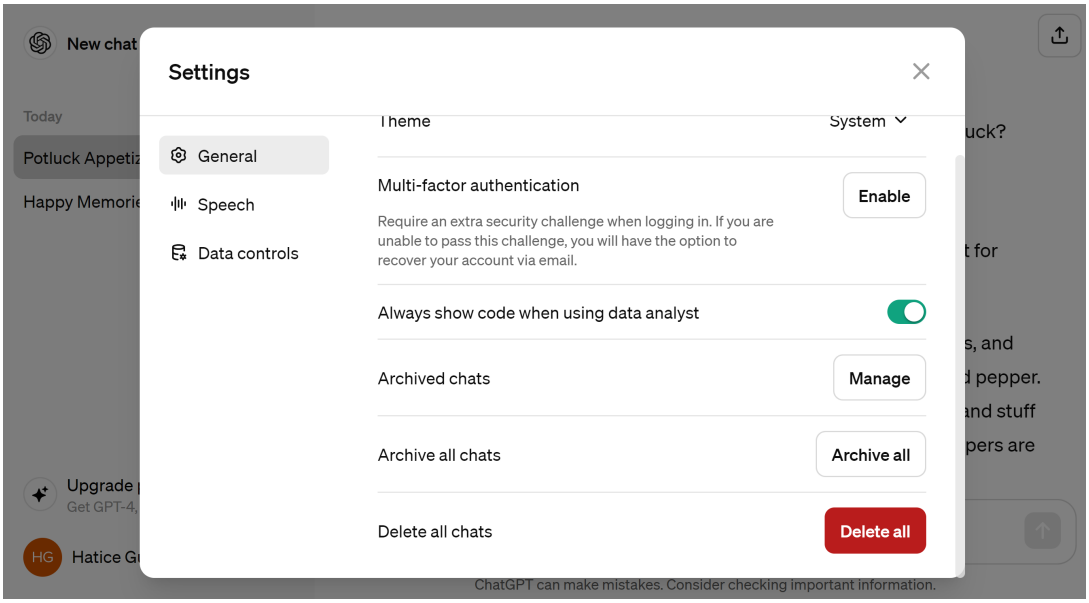


Fig. 3. Clicking on user name and viewing and/or modifying Settings

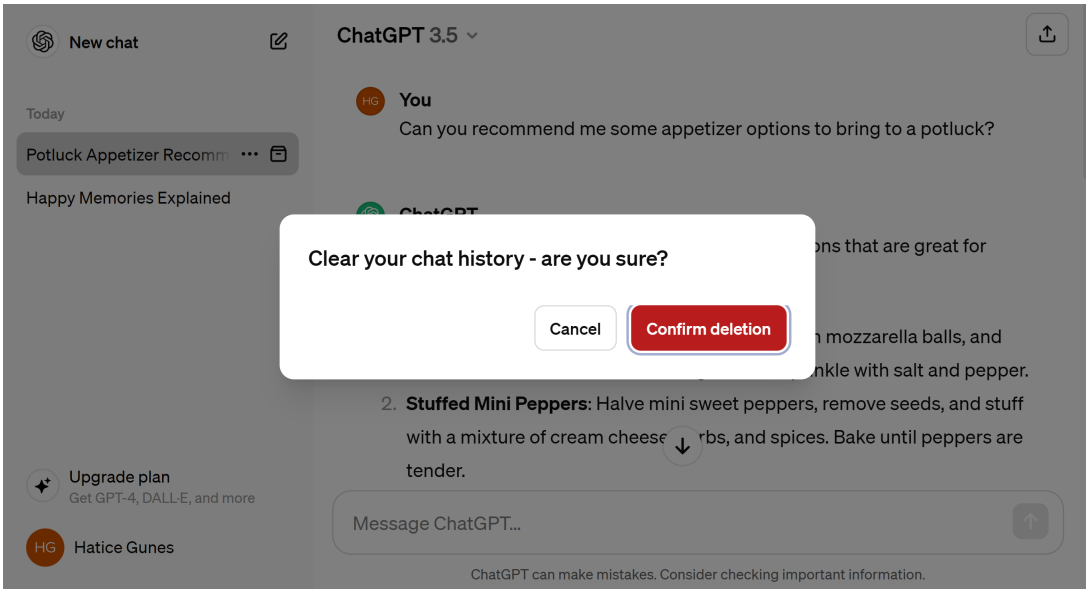


Fig. 4. Clicking on user name. Viewing and/or modifying Settings – Click Delete All Chats

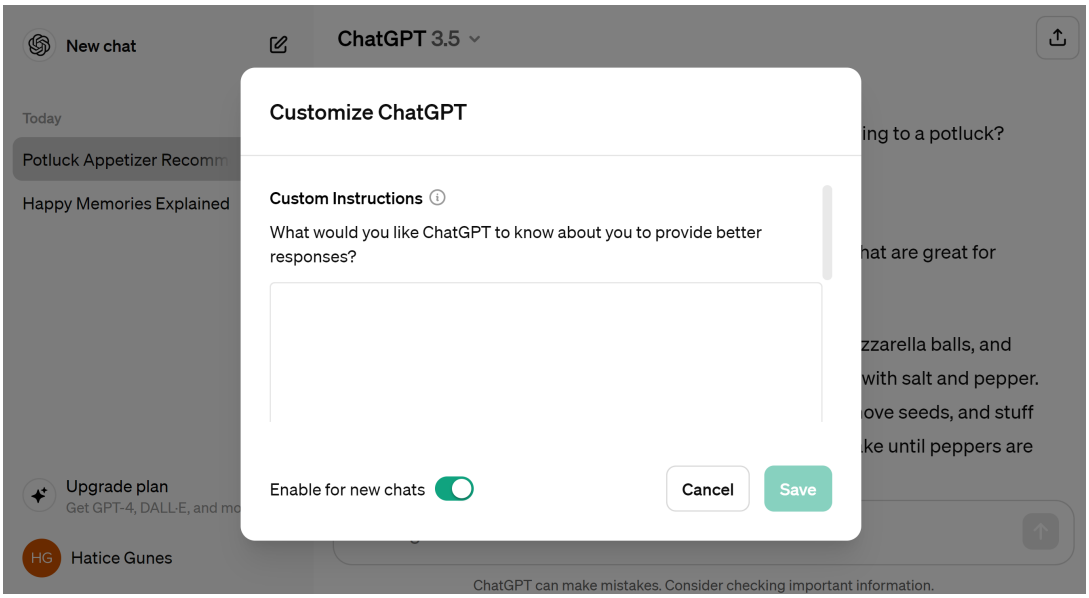


Fig. 5. Clicking on user name and clicking on Custom Instructions

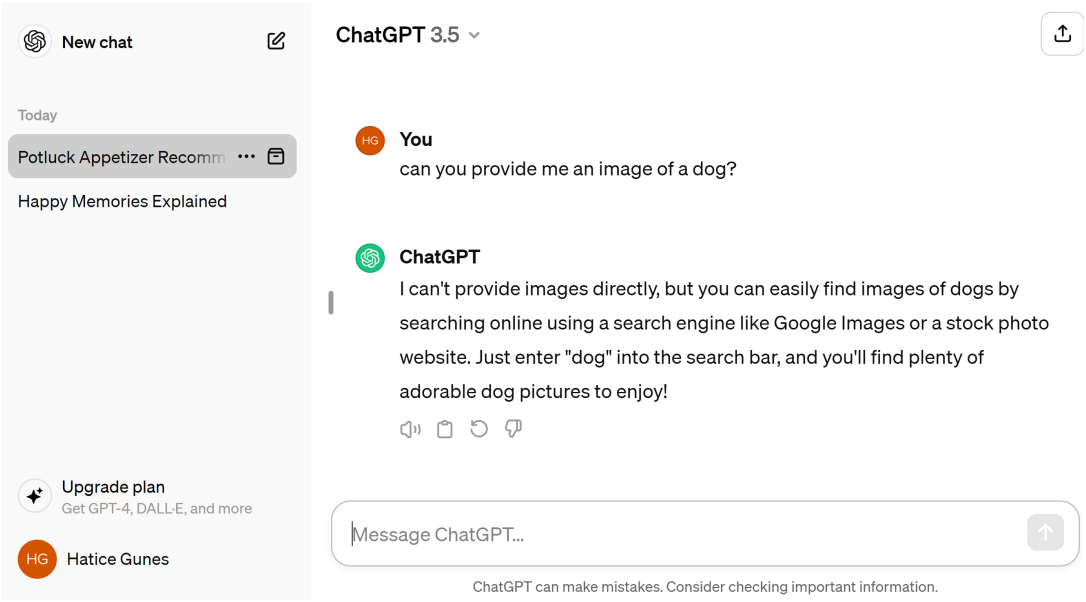


Fig. 6. (starting from the main screen) Asking for an image of a dog

[5 marks]

(*c*) You are working as a member of the Free ChatGPT Unlimited design team, and the management team has asked you to identify whether the current interface needs improvement and why. You decided to conduct interviews with the four groups of stakeholders you have identified in (a). Provide two questions for each stakeholder group to understand whether the current interface needs improvement or not, and why. [8 marks]

(*d*) Describe how you will analyse the interview data gathered in (c), and what your next steps will be to improve the validity and reliability of your findings to support the notion that the current interface needs improvement (or not). [3 marks]

6 Interaction Design

Financial literacy is the ability to understand and effectively use various financial skills, including personal financial management, budgeting, and investing. Low financial literacy threatens the well-being of individuals and families, especially in underserved and low-income communities. Without a solid financial foundation, young people are more susceptible to predatory lending and costly errors in managing debts and expenses that can lead to lifelong financial inequity. Your company has been tasked with creating an Android and iOS app to help young people with financial literacy.

(*a*) Think about the interactive product you are designing, and identify and describe with rationale three key requirements that the application must meet.

[3 marks]

(*b*) Sketch out a design for this app with relevant screens and details, illustrating how it meets the set of requirements you have listed in (a) by providing labels and explanatory captions, and descriptions of the interactive aspects.

[5 marks]

(*c*) Create two user personas for this app and describe how you can use these to guide your design choices. [6 marks]

(*d*) Describe how you will utilise and apply the Gestalt principles to the design you created in (b) to enhance the user experience.

[6 marks]

SECTION D

7 Machine Learning and Real-world Data

A language school for English receives students from different countries and with different skill levels. Before entering, students perform an English test, which decides which class they are assigned to (B1, B2 or I). After studying for a week, students are sometimes reassigned to a different level better reflecting their actual language ability.

(*a*) Professor M is unhappy with this process and the test's ability to predict student level. She suggests that the school should derive students' final level directly by machine learning, based on the students' age, their first language (L1), and how long they studied English before. Several years' data from previous students is available. Describe how the task could be accomplished using a Naive Bayesian Classifier. Apply smoothing if this is appropriate. Give all relevant formulae for parameter estimation and classification. [4 marks]

(*b*) Calculate all relevant probabilities for features age, L1 and experience, for your classifier defined in a), using the following sample of student data. . [6 marks]

ID	Test performance										Student Stats			Level Assignment	
	Total Score	Question 1	2	3	4	5	6	7	8	9	Age	L1	Exper.	Initial	Final
A	4		•	•			•			•	[21-24]	C	[5-7]	B2	B2
B	5		•		•	•	•		•		[13-16]	F	[1-4]	B2	B2
C	1		•								[21-24]	F	[1-4]	B1	I
D	3				•	•				•	[17-20]	C	[5-7]	B1	B2
E	7	•	•	•		•		•	•	•	[13-16]	C	[5-7]	I	B2
F	5	•	•			•	•	•	• •		[21-24]	F	[$\geq$ 8]	B2	I
G	6	•	•		•		•		• •		[17-20]	C	[5-7]	I	I
H	2		•						•		[17-20]	C	[1-4]	B1	B1
I	8	•	•	•	•	•			•	•	[13-16]	F	[$\geq$ 8]	I	I
J	5	•	•			•	•		•		[21-24]	F	[1-4]	B2	B1

(*c*) A new student enters the school whose L1 is "C", who has studied English for 5 years, and who is 18 years old. Which level will this student be assigned to by your classifier from a), as trained in b), and why? [2 marks]

(*d*) Some features influence the prediction of the classifier more than others. How could you use the data available to you to determine the relative relevance of individual features? Describe at least two methods and give a numerical illustration for at least one of your methods, using the above data. [4 marks]

(*e*) The school are now re-thinking how they create classes. Rather than relying on level descriptions such as B1, they want to define classes by grouping incoming students of similar ability into classes of roughly the same size. Describe a method how this could this be achieved. You may use all information in the table above, excluding the information on levels. [4 marks]

8 Machine Learning and Real-world Data

You are interested in predicting snowfall in order to plan for a winter trip with friends. You know that for snowfall to occur, we need to have freezing air temperatures, even if this is not always sufficient.

You decide to model this using a first order hidden Markov model (HMM), with air temperature as the hidden state (Freezing or NotFreezing) and Snowfall, Rain or Dry as the observations. Then you take a look at data from recent months and see:

Month	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun
AirTemp	NF	NF	F	F	F	NF	F	NF	NF
Weather	R	R	S	S	D	R	D	D	D

(*a*) Define and estimate the components of an appropriate HMM for this application, without smoothing. Assume that all hidden states are equally likely to start the sequence. Ignore the start and end states.

[4 marks]

(*b*) Answer the following questions using the model you estimated. Provide the calculations needed to arrive in each answer:

(*i*) What are is most likely combination of air temperature and weather for the next month (July) ?

(*ii*) What is the most likely combination of air temperature and weather in three months from now (September)?

[4 marks]

(*c*) We are in June, and your friends predict that July and August will be dry, followed by rain in September and snow in October. If they are right, what is the most likely sequence of air temperatures for July to October?

[6 marks]

(*d*) The current model you have estimated ignores the time of the year we are in, e.g., the probability of transitioning from NotFreezing to Freezing is independent of whether we are in summer or in winter. Build a better model by taking into account the time of the year, by changing the definition of hidden states. Describe any transformations necessary to the data and estimate the parameters of the new model.

[6 marks]

9 Machine Learning and Real-world Data

You are a football academy manager and you want to design a classifier for deciding which players to recruit to your academy. For this purpose you gather recent data (2020 and onwards) on players about to finish high school. You construct the following table, where each row represents a potential recruit. The column `Success` indicates whether the player was considered a successful recruit or not; it is the label we are trying to predict.

Success	Goals	Position	Gender
Y	Many	Attack	M
N	None	Goalkeeper	F
Y	Few	Defender	M
Y	Few	Attack	M
N	None	Defender	M
N	Few	Defender	F

You want to develop a model to help you in deciding whether you should recruit a player based on their features. You decide to use a Naive Bayes Classifier.

(a) Define and estimate the parameters of the Naive Bayes Classifier. [4 marks]

(b) There is a suspicion that the model you estimated is biased. If a model is biased, it treats certain groups of the population unfairly, i.e. it never predicts their members to be successful. Identify two such groups and demonstrate the problem by constructing appropriate test instances. [4 marks]

(c) Explain which property of the classifier you developed in part (b) enables you to make this judgement based on the parameters you have estimated. Give one reason why it is a useful property for the model to have and one reason why this is a problematic property. [2 marks]

(d) You are given more data from an earlier time period, pre-2020. How would you incorporate it in your experimental setup for the classifier you developed in part (a)? [4 marks]

(e) The model you developed predicts success or not based on a snapshot of the potential recruits at a single point in time. However you are recruiting for an academy, so you want to take into account the trajectory of the players over a number of years. Propose a modelling solution for this. [6 marks]

END OF PAPER